Noname manuscript No.
(will be inserted by the editor)

Genetic and individual assignment of tetraploid green sturgeon with SNP assay data

Eric C. Anderson · Thomas C. Ng · Eric D. Crandall · John Carlos Garza

Received: date / Accepted: date

Abstract Polyploid organisms pose substantial obstacles to genetic analysis, as molecular assay data are usually difficult to evaluate in a Mendelian framework. Green sturgeon (*Acipenser medirostris*) is a tetraploid species and is facing significant conservation challenges, including bycatch in ocean fisheries. We present here novel molecular genetic assays and analytical methodology for green sturgeon that allow discrimination of fish from the two visually indistinguishable distinct population segments (DPSs), and also provide individual-specific genetic tags. We show how the relative fluorescence intensity data from a standard quantitative PCR assay, designed for a biallelic single nucleotide polymorphism, can be grouped into "genotype categories" using standard analytical software and post-processing manipulation. We then show how these genotype category data can be used to discriminate green sturgeon from the southern DPS, which is protected under the US Endangered Species Act, and the northern DPS, which is not. We also show how these data can be used to reliably identify individual green sturgeon, and can therefore be used in capture/recapture analyses. Both types of identification are extremely accurate even when fewer than half of the assays are successfully called. We then apply these new techniques to show that proportions of the two green sturgeon DPSs are extremely different in the two major fishery areas where they are encountered as bycatch. While these assays and methods do not provide data that can be used in pedigree-based analyses, they are an important advance in the application of genetic analysis to conservation and management of polyploid organisms.

E.C. Anderson
Fisheries Ecology Division, Southwest Fisheries Science Center, National Marine Fisheries Service *and* University of California, Santa Cruz, 110 McAllister Way, Santa Cruz, CA 95060, USA. E-mail: eric.anderson@noaa.gov

T.C. Ng
Department of Biomolecular Engineering, University of California *and* Southwest Fisheries Science Center, National Marine Fisheries Service, 110 McAllister Way, Santa Cruz, CA 95060, USA

E.D. Crandall
University of California *and* Southwest Fisheries Science Center, National Marine Fisheries Service, 110 McAllister Way, Santa Cruz, CA 95060, USA
Present address: of E.D. Crandall
Division of Science and Environmental Policy, California State University, Monterey Bay, 100 Campus Center, Seaside, CA 93955, USA

J.C. Garza
Fisheries Ecology Division, Southwest Fisheries Science Center, National Marine Fisheries Service *and* University of California, Santa Cruz, 110 McAllister Way, Santa Cruz, CA 95060, USA.

1 Introduction

Polyploidy, the presence of more than two copies of the genome in an individual, is relatively common in plants and invertebrates, but uncommon in vertebrates (Mable 2004). Sturgeon, fishes in the family *Acipenseridae*, have a number of species that are polyploid and the mosaic of ploidy in the group suggests a polyploid ancestor, with subsequent rediploidization in some species and further genome duplication in others (Ludwig et al. 2001), although high homology between genome copies in some species contradicts the polyploid ancestor hypothesis (Havelka et al. 2011). While the causes of this unusually high prevalence of polyploidy in sturgeon remain to be elucidated, its consequences for molecular population genetics are clear: the presence of more than two copies of each gene region greatly complicates use

of molecular genetic data in all methods that make an assumption of Mendelian segregation.

We describe here novel molecular genetic markers and methodology for the study of one such polyploid fish species, green sturgeon (*Acipenser medirostris*). Green sturgeon are anadromous fish inhabiting the northeastern Pacific and its tributaries. Spawning activity is known to occur primarily in the Sacramento, Klamath and Rogue river basins (Adams et al. 2007), although adult green sturgeon enter other rivers and evidence of spawning activity has recently been observed in the Columbia River (Schreier et al. 2016) and in the Eel River of California (J. Strange and S. Kullman, pers. comm.)

Previous status reviews (Adams et al. 2007; NMFS 2015) have determined that there are two Distinct Population Segments (DPSs) of green sturgeon. The Southern Green Sturgeon DPS includes fish originating from the Sacramento River, and the Northern Green Sturgeon DPS includes fish originating from the Klamath and Rogue rivers. Previous evaluations (Adams et al. 2007; NMFS 2015) have determined that the Southern DPS is threatened with risk of extinction and it has been listed for protection as such under the US Endangered Species Act (ESA). The Northern DPS is a National Marine Fisheries Service (NMFS) Species of Concern.

Effective conservation and management of green sturgeon requires identifying fish from these visually indistinguishable DPSs when they are encountered in ocean fisheries and elsewhere and we describe the development of genetic assays and analytical methodology that allow such identification. Specifically, we have used next generation (high-throughput) DNA sequencing to identify variable sites (variants) in the green sturgeon genome and used this information to develop a panel of molecular assays that target those sites and can be used to characterize these variants in individual fish. The initial application goal for these genetic markers was to use them in identifying individual fish to their DPS of origin. The secondary objective was to evaluate their ability to identify specific individuals that have been sampled more than once (i.e., as DNA "fingerprints").

Green sturgeon are not only tetraploid but they appear to have experienced a recent duplication of their entire genome, with very high homology of the different genome copies (Israel et al. 2009). As such, they are not amenable to traditional population genetic evaluation using the methods developed for calling genotypes in diploid species, since designating alleles to genome copies is nearly impossible. We have therefore developed a genotyping system and workflow to allow calling of such variants in tetraploid species. This methodology uses quantitative PCR assay data but does not call the individual alleles and the dosage (number of copies) carried by each tetraploid individual, but, rather, places individuals into "genotype categories." The inheritance properties of such genotype categories are not directly considered in the statistical analysis, but the data are still suitable for both the assignment of individuals to DPS and the identification of samples from the same individual.

We describe below how these new markers and methods uncover a substantial amount of genetic differentiation between the two green sturgeon DPSs, and that the panel of SNP markers and the statistical procedures we have developed provide a robust, cost-effective, and extremely accurate means of identifying the DPS of origin for individual green sturgeon and of identifying green sturgeon that have been sampled multiple times. We then use genetic assay data from green sturgeon caught as bycatch in ocean fisheries off the west coast of North America to start to elucidate population-specific patterns of ocean distribution. We also identify the DPS of origin of fish that apparently represent a recolonization of part of the species historic reproductive range. The analytical methodology we describe, assigning genotype categories rather than calling specific genotypes that have an assumption of Mendelian segregation, has broad relevance to the study of polyploid species and will allow biological inference regarding identity at the population and individual level to be derived from standard SNP genotyping assay data.

2 Methods

2.1 Genetic Samples

We analyzed samples of three general categories. The first is from fish sampled in two rivers known to have well-established spawning populations of green sturgeon—the Sacramento River and the Klamath River. These samples were used to construct the "reference" or "baseline" dataset. All of these samples were taken from juvenile fish or from eggs collected on egg mats, either of which must have been spawned (originated) in the river from which they were sampled. We refer to these as "egg/juvenile reference" samples. There were also a few samples taken from adult or sub-adult fish in the Sacramento River that may be non-spawning visitors, originating in a different river, and these are referred to as "non-juvenile reference" samples. Samples taken from the ocean or from rivers without a persistent spawning population are "non-reference" samples. These include green sturgeon caught as bycatch in ocean fisheries, and five fish sampled from the Eel River, CA. See Table 1 for a summary of sample types and sample sizes. Tissue

Table 1 Summary of genetic samples used in this study.

Location	Category	Life Stage	Group Short Name	n	Collection Years
Sacramento River	reference	egg	reSac	66	2011–2012
Sacramento River	reference	juvenile	rjSac	72	2012–2013
Sacramento River	reference	non-juvenile	rnSac	12	2014
Klamath River	reference	juvenile	rjKla	21	2006
Eel River	non-reference	non-juvenile	nnEel	5	2015
Bycatch	non-reference	non-juvenile	nnByc	190	2008–2014

samples were of various types, including fin clips, either dried or stored in ethanol, and whole eggs or whole juvenile specimens stored in ethanol. Freshwater samples were obtained primarily during the late spring and early summer, while bycatch samples were obtained from all months of the year except September and December.

2.2 SNP Discovery and Assay Development

To identify SNPs to be made into assays, we first collected DNA sequence information from eight green sturgeon, four juveniles from the Klamath River (Northern DPS) and four juveniles from the Sacramento River (Southern DPS). These are referred to as the "ascertainment" samples. We used a double-digest restriction-site associated DNA protocol (ddRAD; Peterson et al. 2012), which uses restriction enzyme digestion and DNA fragment size selection to reduce the fraction of the genome that is analyzed, ensuring that a sufficiently small portion is sequenced to provide adequate numbers of DNA sequences (read depth) from the eight different individuals. Following the Peterson et al. (2012) protocol, genomic DNA was digested with EcoR1 and Sph1 restriction enzymes, cleaned with Ampure XP beads (Beckman Coulter) and quantified using a Qubit Fluorometer (ThermoFisher Scientific). Enzyme-specific adapters containing sample-specific barcodes were then ligated, and the individual libraries were pooled in equimolar concentration. Pooled libraries were size selected to a target length of 500 base pairs using a Pippin Prep (Sage Science). Size-selected libraries were again quantified and measured for length distribution using the Qubit and a Bioanalyzer (Agilent). We used streptavidin beads to remove adapters that had self-ligated and then performed 12 cycles of PCR using primers that added Illumina flowcell annealing sequences to the library. Following another cleanup with Ampure beads, the library was qPCR quantified with Kapa SYBR Fast qPCR master mix and then sequenced on an Illumina MiSeq instrument, yielding 21.24 million sequence reads, each of 300 bp in length. We used Stacks software v 1.12 (Catchen et al. 2011) under the following parameter settings—$m = 4$, $M = 2$, $n = 2$—to assemble the sequences into 84,747 putatively unique genetic loci. Within those loci, we identified 21,218 SNPs with minor allele frequencies greater than 0.25 in the sample of eight fish. At 1,118 of the SNPs, the four Klamath fish had a single variant that was different from the single variant carried by the four Sacramento fish. Of these 1,118 SNPs, we chose 96 with the highest read depths and quality scores for SNP Type™ genotyping assay design (Fluidigm Co., South San Francisco, CA). Although these SNPs appeared to be fixed for alternate variants in the ascertainment samples from the two DPSs, many of them were polymorphic in the large set of reference samples (see Results).

The 96 assays were then validated on 96.96 SNP Genotyping Arrays (Fluidigm) which use nanofluidic technology to simultaneously genotype all 9,216 combinations of 96 assays and 96 individuals on a single array (hereafter called "chip"). These assays contain two gene probes that are specific to the two variants uncovered in the next generation sequencing effort. These probes carry fluorescent dyes that measure the relative abundance of the two variants and report them as relative fluorescence intensities in a two-dimensional coordinate space. Each variant at a SNP therefore corresponds to a fluorescent dye and the genotype of a diploid individual can be called from the intensity of fluorescence for each of those dyes. Fluidigm provides software that allows genotype calling of diploid individuals, but the software is not designed for the calling of tetraploid genotypes.

2.3 Developing Calling Methodology

Apparent in our data were two difficult features. First, data from most of the green sturgeon SNP assays did not consist of five clearly callable fluorescence clusters (genotypes), as expected for tetraploids; rather, many of the clusters in the fluorescence intensity data were indistinct or overlapping. Second, we encountered consid-

Table 2 Four scoring-development chips

(a) Number of samples from each location

Chip	Sacramento	Klamath	Bycatch
1	0	17	77
2	54	17	23
3	78	8	8
4	37	2	55

(b) Number of samples shared between chips

Chip	1	2	3	4
1	94	13	6	9
2		94	8	4
3			94	32
4				94

erable chip-to-chip variation in fluorescence intensity. To deal with the indistinct clustering, it suffices to put an individual into a particular *genotype category* without reference to the underlying alleles. In this scheme, individuals with different *genotypes* may be put in the same genotype *category* if the genotypes are difficult to resolve on the basis of their fluorescence intensities. To address the chip-to-chip variation, each chip includes samples from a number of individuals that have been previously analyzed, so that corresponding clusters between the chips can be identified and the corresponding fluorescence intensities used to calibrate calling across chips (Figure 1).

We first used four chips to decide which loci were consistently callable (*call-development chips*) and, at each locus, the number of genotype categories that could be resolved. Some individuals were genotyped on multiple chips to evaluate chip-to-chip variation (Table 2). We then created a series of graphs (one for each locus) that plotted the raw fluorescence intensities from the Fluidigm chip reader (Figure 2). The *raw* intensities were not transformed based on the observed intensity of the two no-template controls (NTCs) that are included on every chip. We used the raw intensities because there was less chip-to-chip variability in them compared to the default NTC-normalized values from the Fluidigm software. On these graphs, results from each chip are in a different color and points for the same sample genotyped on different chips are connected with line segments. The plots of four exemplar loci with 2–5 genotype categories regarded as callable are shown (Figure 2). Similar plots for all of the loci are in the Supplement 1. In general, we preferred to merge nearby clusters into a single genotype category, ensuring that genotype categories could be called reliably, rather than maintaining more genotype categories, some of which could be subject to high rates of categorization error. For example, at locus ame_21745 (Figure 2*a*) there are three clusters apparent, but two are rather close together and could be difficult to resolve in the face of substantial chip-to-chip variation. Accordingly, we choose to call only two genotype categories for this locus.

For subsequent chips we followed a simple routine. First, we included between 16 and 24 individuals from the call-development chips on each new chip. Second, we plotted the raw fluorescence intensities of the new chip along with the raw intensities of the four call-development chips, including line segments connecting the results from the 16 to 24 individuals on the old and the new chips to verify the correspondence between clusters. We then analyzed the new chip using the Fluidigm SNP Genotyping software which allows the user to call up to three clusters of fluorescence intensity. Because this software allows genotypes to be called into only three categories (as expected for a SNP locus with two alleles in a diploid species), we called genotypes into these three allowed categories and used a custom R script (R Core Team 2015) to convert those calls into up to five categories by using a combination of the genotype calls and the fluorescence intensities. As an example, when we designate a genotype in the Fluidigm software as a '1'—a homozygote that typically clusters in the lower right part of the X-Y plot of fluorescent intensity—but its intensities are such that it falls in the upper left part of the X-Y plot, then, we convert that call to genotype category '4'. If the point fell in the lower right part of the plot, then it would have remained in genotype category '1'. To ensure that this combined calling procedure, with initial Fluidigm software analysis followed by post-processing adjustment to resolve the five genotype categories, can be done easily by others, our entire workflow is available and documented at `https://github.com/ngthomas/sturgeon_fluidigm` and archived at Dryad XXX-XXX.

2.4 Population structure, population assignment, and individual assignment

Our calling method does not resolve the exact genotype of each individual. That is, the precise number of copies of each different allele within an individual is not obtained. Accordingly, these data are not appropriate for use with analytical methods that assume Mendelian segregation or information about the individual, constituent alleles of each individual's genotype. However, it is still valid to apply a number of statistical analyses, including the model-based clustering method in the program *structure* by assuming haploid inheritance for each genotype category, and using the model without admixture. In this case, the software implements a finite

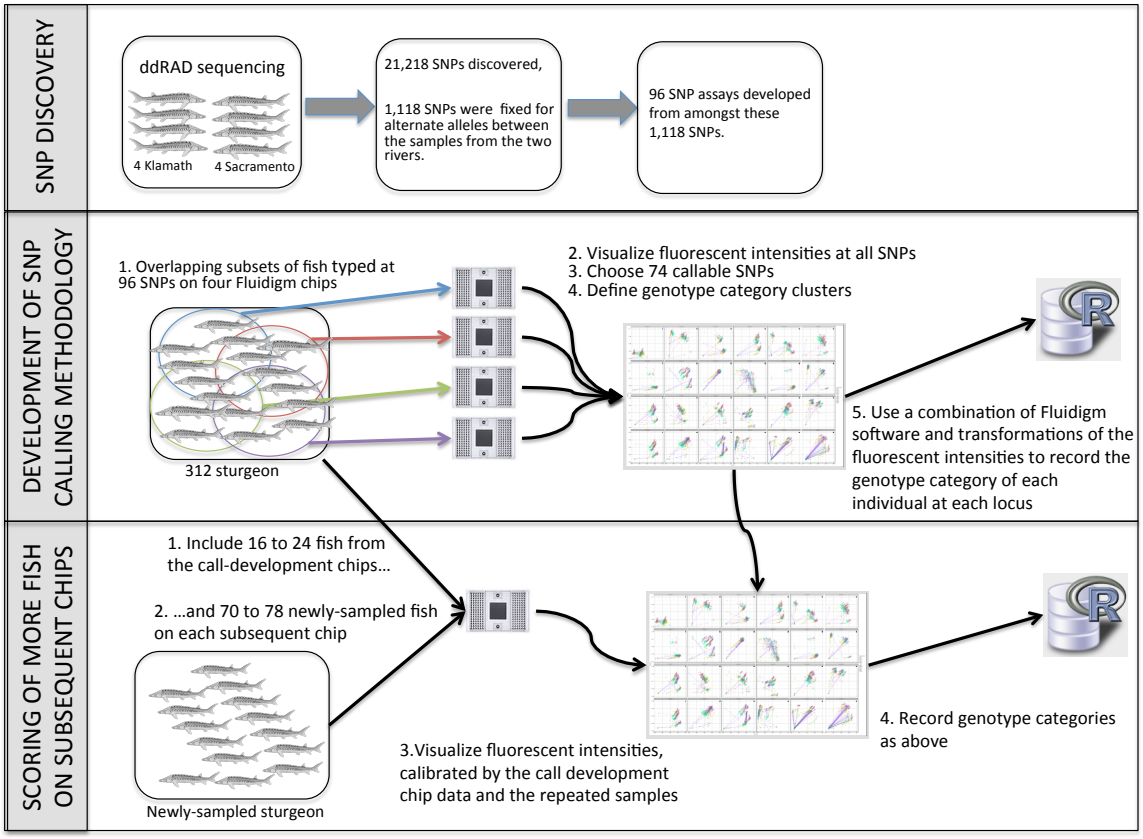

Fig. 1 Graphic depiction of workflows. SNP discovery is done only once. Development of SNP calling methodology is undertaken only once. The steps for scoring of more fish on subsequent chips occur once for each additional chip that is run.

mixture model in which the component-specific parameters are the frequencies of different genotype categories at each locus, and an individual's genotype category is assumed drawn, independently at each locus, from these frequencies. Such a model does not require any assumptions about constituent alleles or the mode of their inheritance.

It is important to understand that this approach with *structure* is quite different than coding the individual alleles in polyploids as dominant markers by including the presence/absence of each allele as a binary phenotype. The dominant-marker approximation has been useful in microsatellite surveys of polyploid seaweed and sturgeon (Andreakis et al. 2009; Schreier et al. 2013); however with SNP assay data, clustering genotypes under a haploid model makes fewer assumptions about inheritance of the markers, and, as such, makes fewer approximations when populations are not admixed.

We included all individuals from which we had successfully obtained calls at 60 or more assays in an analysis with *structure* at K (the number of genetic groups or populations) equal to one, two, and three. For each value of K, we ran *structure* 18 times with a burn-in of 5,000 sweeps and 20,000 sweeps of data collection. We used no prior information about the origin of the reference samples. Our aim was to evaluate the genetic clustering of individuals from the "reference" sample collections and whether they consistently grouped into two clusters, coincident with the DPS categorization.

Similarly, population assignment or, as it is often called in fisheries, "genetic stock identification" (GSI), can be carried out under a haploid model, using our definition of genotype categories. The model underlying GSI is nearly identical to the "without admixture" model of *structure*, except that 1) *structure* assumes that the proportion of individuals from each component is equal, while in GSI that proportion is estimated, and 2) *structure* can perform unsupervised clustering, while for GSI, known samples from each DPS are used as "baseline" samples. We used the reference samples from the Sacramento and Klamath rivers as our baseline samples from the Southern and Northern DPSs, respectively.

Our reference samples did not include any fish from the Rogue River, the only other river with consistent re-

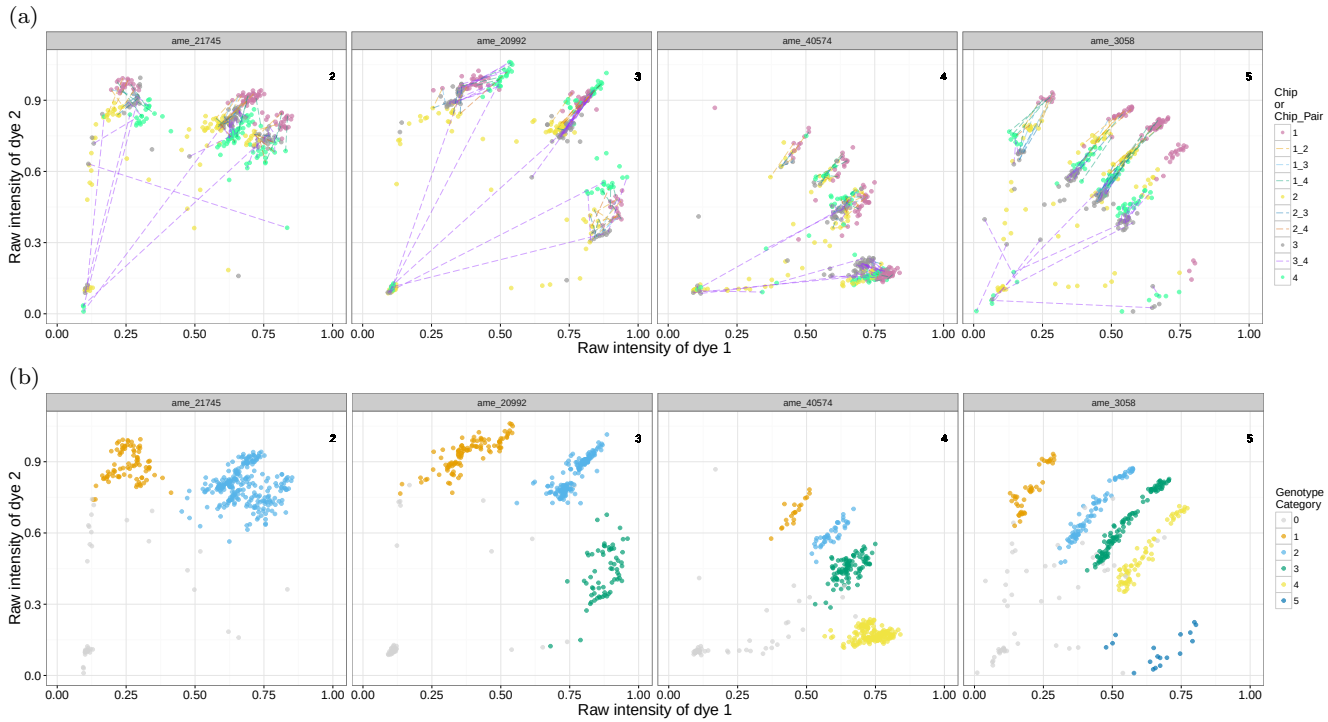

Fig. 2 Raw SNP Type™ fluorescence intensities at four loci typed upon the four scoring-development chips. Each panel shows the results at four chips for one locus. The numeral in the upper right gives the number of genotype categories that are scored at the locus. Each point represents an individual's fluorescence intensities at two dyes. (*a*) Individuals are colored according to chip and dashed line segments connect the intensities of individuals typed on different chips. (*b*) Individuals are colored according to the genotype category to which they have been scored. Category 0 (colored light gray) is a "no call", *i.e.*, the individual is recorded as having missing data at the locus. Supplement 1 contains plots like these for all 96 assays.

production by green sturgeon. However, a previous microsatellite study (Israel et al. 2009) showed that green sturgeon in the Klamath River and the Rogue River are not greatly differentiated ($F_{\mathrm{ST}} \approx 0.01$) relative to the differentiation between the Northern and Southern DPSs ($F_{\mathrm{ST}} \approx 0.05$–$0.08$). This suggests that the absence of Rogue River in our baseline should not greatly reduce our ability to assign fish to the Northern DPS, even if they originate from the Rogue River.

We assessed accuracy of self-assignment of the baseline samples using a leave-one-out procedure, and then we performed GSI on the remaining samples (fishery bycatch and Eel River samples), using the expectation-maximization algorithm (Dempster et al. 1977) to maximize the likelihood. All analyses were done using the software *gsi_sim* (Anderson et al. 2008; Israel et al. 2009), which returns an estimate of the posterior probability that each fish belongs to the Southern or Northern DPS.

We also investigated whether the observed genotype categories can be used to identify the same individual when it is sampled more than once. For this purpose, we developed a straightforward log-likelihood ratio statistic to discriminate genotypes from samples taken from the same individual green sturgeon and those taken from different individuals. We compared the distribution of the statistic in both simulated and observed genetic assay data from pairs of samples known to be from the same fish, both from DNA samples subjected to genotyping more than once and from independent DNA extractions from tissue taken from the same individual.

To derive our log-likelihood ratio statistic, let $\boldsymbol{p}_\ell$ be the frequencies of the genotype categories at locus ℓ in the DPS from which a pair of individuals originate. For example, if four categories are scored at locus ℓ, then $\boldsymbol{p} = (p_{\ell,1}, \ldots, p_{\ell,4})$, with $\sum_{k=1}^{4} p_{\ell,k} = 1$. Assume that at locus ℓ a fraction ϵ_ℓ of the samples is expected to be subject to genotyping error. If a sample is subject to genotyping error, then its observed genotype category is drawn from $\boldsymbol{p}_\ell$, independently of its true, unobserved genotype category. Though this is a somewhat unrealistic genotyping error model, it is mathematically convenient and adequate for our purposes. Under this assumption, the marginal probability that a sample, i, is observed to have genotype category g at locus ℓ is

$$P(y_\ell^{(i)} = g) = (1 - \epsilon)p_{\ell,g} + \epsilon p_{\ell,g} = p_{\ell,g}.$$

If another sample j that is unrelated to sample i—we shall say $K_{ij} = \mathrm{U}$—is typed and found to have genotype

category h (which may be the same category as g) then the joint probability is simply

$$P(y_\ell^{(i)} = g, y_\ell^{(i)} = h | K_{ij} = \mathrm{U}) = P(y_\ell^{(i)} = g) P(y_\ell^{(j)} = h) = p_{\ell,g} p_{\ell,h} \quad (1)$$

If i and j are two samples from the same individual ($K_{ij} = \mathrm{S}$), then the joint probability of the observed genotypes can be written as a sum over the four possible cases: 1) neither sample suffered a genotyping error; 2) i suffered an error, but not j; 3) j suffered an error, but not i; or 4) both i and j suffered an error:

$$\begin{aligned} &P(y_\ell^{(i)} = g, y_\ell^{(i)} = h | K_{ij} = \mathrm{S}) = \\ &(1-\epsilon_\ell)^2 p_{\ell,g} \delta(g = h) \\ &+ \epsilon_\ell p_{\ell,g} \times (1-\epsilon_\ell) p_{\ell,h} \\ &+ (1-\epsilon_\ell) p_{\ell,g} \times \epsilon_\ell p_{\ell,h} \\ &+ \epsilon_\ell^2 p_{\ell,g} p_{\ell,h}, \end{aligned}$$

where $\delta(g = h)$ is 1 when $g = h$ and 0 otherwise. This can be written more compactly as

$$\begin{aligned} &P(y_\ell^{(i)} = g, y_\ell^{(i)} = h | K_{ij} = \mathrm{S}) = \\ &(1-\epsilon_\ell)^2 p_{\ell,g} \delta(g = h) + (2\epsilon_\ell - \epsilon_\ell^2) p_{\ell,g} p_{\ell,h}. \end{aligned} \quad (2)$$

A log-likelihood ratio statistic that can be used to identify pairs of samples i and j from the same individual is then

$$\Lambda_{ij} = \log \frac{P(y_\ell^{(i)} = g, y_\ell^{(i)} = h | K_{ij} = \mathrm{S})}{P(y_\ell^{(i)} = g, y_\ell^{(i)} = h | K_{ij} = \mathrm{U})}. \quad (3)$$

To approximate the distribution of Λ_{ij} under $K_{ij} = \mathrm{S}$ and $K_{ij} = \mathrm{U}$, we conducted Monte Carlo simulations using values of $\boldsymbol{p}$ for the Northern and Southern DPSs obtained by counting genotyping categories scored on individuals assigned by *structure* to the Northern DPS and Southern DPS clusters. For the simulation of pairs and for the calculation of $P(y_\ell^{(i)} = g, y_\ell^{(i)} = h | K_{ij} = \mathrm{S})$ for each simulated pair, we assumed a conservatively high genotyping error rate of $\epsilon_\ell = 0.05$ for every locus. 10^5 genotype pairs were simulated both for $K_{ij} = \mathrm{S}$ and $K_{ij} = \mathrm{U}$. Simulations were first done for the case in which every pair was scored at all of the SNPs developed in our panel. Then, to investigate the expected distributions when pairs shared fewer scored loci due to missing data (*i.e.*, loci that were not scored), we successively deleted loci randomly, without replacement, according to the observed locus-specific missing data rates in our data set, from each simulated pair, and recomputed the Λ_{ij} values. We computed Λ_{ij} for all pairs ij of fish from amongst those assigned by *structure* to the Northern DPS or the Southern DPS. Both the simulations and the pairwise comparisons of real fish were done using a custom-written script in R (R Core Team 2015).

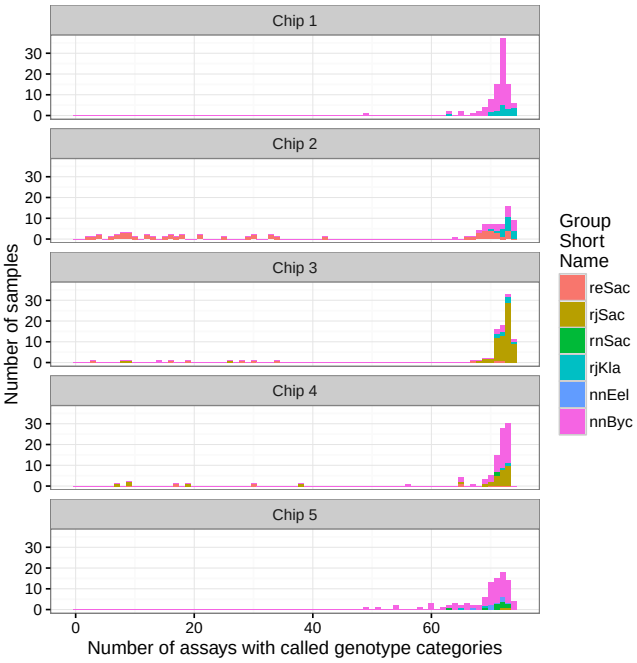

Fig. 3 Histogram showing the number of individuals successfully scored for $0, \ldots, 74$ assays, on each of the chips. Color denotes type of sample; Group Short Names as in Table 1.

3 Results

3.1 Loci scored

Of the 96 SNP assays developed, 74 of them could be reliably called into genotype categories (Supplement 2). Of these, 39 were loci with two categories, 29 with three, 3 with four, and 3 with five. Of the 470 total DNA samples analyzed with these SNP assays, 413 (corresponding to 319 distinct green sturgeon, since some were genotyped more than once) yielded called genotype categories for at least 60 of the 74 loci. Inspection of the distribution of the number of called loci revealed that only the egg samples from the Sacramento River consistently failed to genotype well (<60 called assays; Figure 3).

3.2 Analysis with *structure*

The 319 individuals with 60 or more called loci were included in the clustering analysis with *structure*. The posterior probability that each individual originated from one of the K clusters is given in Figure 4. The results show unambiguously that the model for population structure that is best supported by the data is one with $K = 2$ genetic groups. At $K = 2$, individuals were either allocated entirely to one population or another, and the same result was obtained during every *struc-*

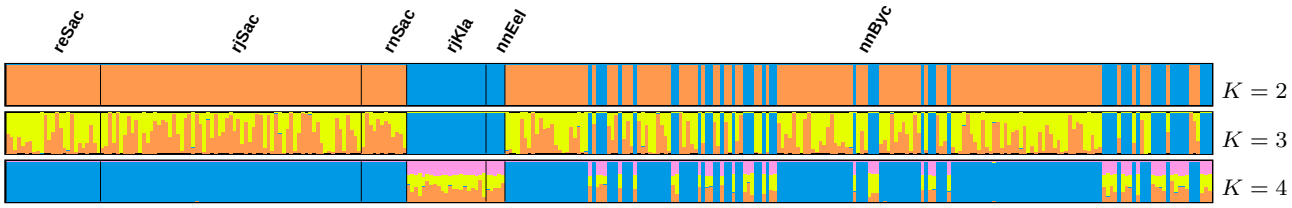

Fig. 4 Representative *distruct* (Rosenberg 2004) plots of *structure* estimates of posterior probabilities of cluster membership for 319 individuals with ≥ 60 loci. Sampling location codes are defined in Table 1. Each individual is represented by a vertical bar and the amount of different colors in those bars represents the posterior probability of cluster membership. For example, with $K = 2$ each bar is either completely orange (Northern DPS cluster) or completely blue (Southern DPS cluster), indicating that individuals can be assigned to those clusters with effectively no uncertainty.

ture run, regardless of the random seed used. For $K = 3$ and $K = 4$, there was more variability, indicating that a model with more than two genetic groups of green sturgeon was not supported by the genetic data. By noting the groups to which the reference samples are assigned, it is clear that these two groups coincide perfectly with the Northern and Southern DPSs of green sturgeon. In addition, it is clear that the fish sampled from the Eel River are genetically similar to those from the Klamath River in the Northern DPS.

3.3 Population assignment with *gsi_sim*

For performing GSI, we used data from all green sturgeon individuals, regardless of the number of successfully called loci. The known-origin reference samples from the Southern DPS (Sacramento River) were from 150 individuals, many of which were egg samples that did not genotype well. The known-origin reference samples from the Northern DPS (Klamath River) consisted of 21 fish. Consistent with the *structure* results, self-assignment of fish in the baseline was very accurate. All of the fish from the Klamath were self-assigned to the Northern DPS with posterior probabilities > 0.999. Likewise, all the Sacramento fish were self-assigned to the Southern DPS except a single fish that was called at only four loci. All of the fish in the bycatch were assigned to the Northern or Southern DPS with posterior probabilities > 0.999. The poor genotyping success with the eggs from the Sacramento River allowed us to investigate the effect of the number of called genotypes on the confidence in assignments (Figure 5). This revealed that a correct assignment to DPS can be made with high confidence even if as few as 10 SNPs were successfully called, highlighting the genetic divergence between the two DPSs and indicating that the SNP markers will be capable of correctly assigning green sturgeon to DPS even with degraded tissues or otherwise low-quality DNA.

The green sturgeon bycatch consisted of fish that are from both the Northern and Southern DPSs, with a greater proportion from the Southern DPS. The bycatch was all from two fishery areas: Gulf of the Farallones, near San Francisco, CA, and around the Columbia River plume, from Tillamook, OR to Grays Harbor, WA. Of the $n = 104$ fish sampled in the southern area, all but one were from the Southern DPS. By contrast, of the $n = 86$ fish sampled in the northern area, 39 were from the Southern DPS and 47 were from the Northern DPS (Figure 6).

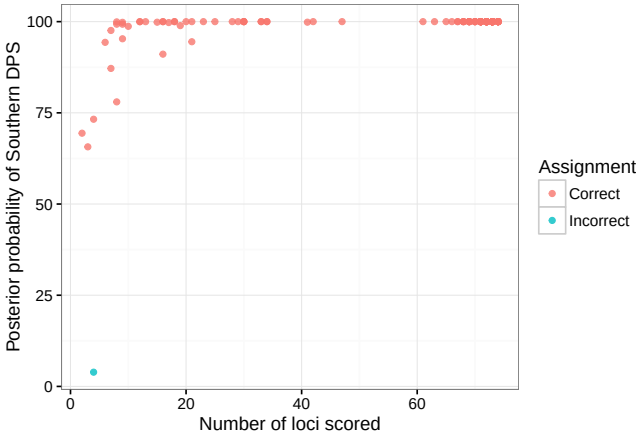

Fig. 5 Relationship between number of loci score and the posterior probabilities of self-assignments (using a leave-one-out procedure) of Sacramento River fish and eggs to the Southern DPS.

3.4 Individual identification

Both the simulated (Fig. 7*a*) and observed (Fig. 7*b*) data revealed completely non-overlapping distributions of the likelihood ratio statistic for pairs of samples from unique green sturgeon and pairs from the same individual. This is true even when there is only data from 40 of the SNP assays shared between the two genotypes. This comparison demonstrated unambiguous identification of samples from the same individual.

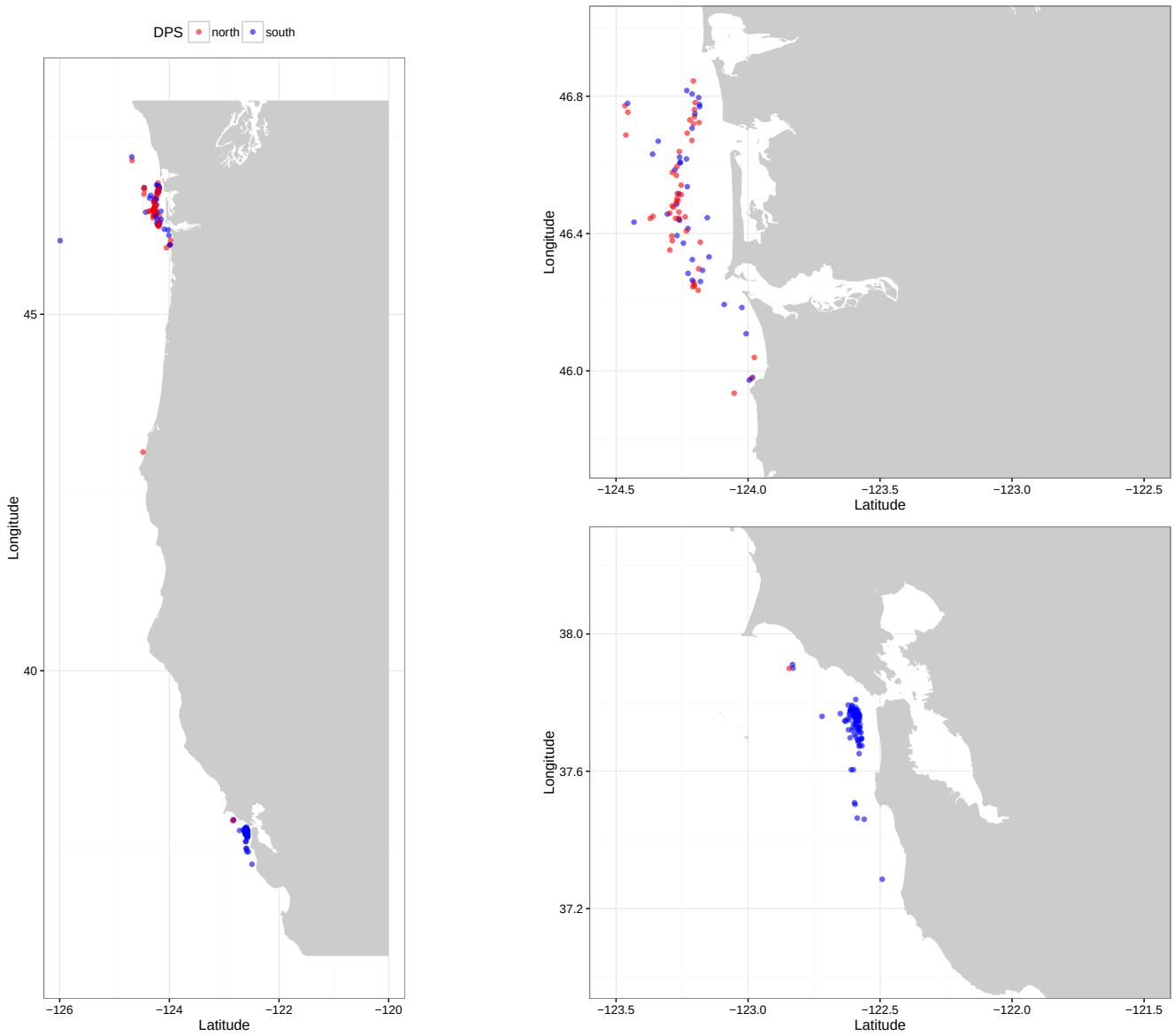

Fig. 6 Location and DPS-origin of fishery bycatch samples. Left panel is a coastwide view. The two right panels show the Columbia Plume region (top) and the Gulf of the Farallones (bottom). Each point represents one green sturgeon. Red = Northern DPS. Blue = Southern DPS. Locations shown have been randomly jittered by up to 0.02° of latitude and longitude.

4 Discussion

Population genetic analysis of polyploid species has required development of an array of novel methods to overcome the obstacles posed by the difficulty in assigning gene copies to their respective loci. We report here the development of a panel of 74 SNP genotyping assays that target variable regions of the polyploid green sturgeon genome and a novel analytical method that skirts the inability to determine Mendelian relationships between gene copies by grouping them into "genotype categories." These assays and this calling methodology provide informative data for several applications, including standard genetic stock identification (GSI) and individual (re)identification.

We show that GSI with these data can accurately assign all individual green sturgeon to either the Northern or the Southern Distinct Population Segment (DPS) with extremely high confidence, even with a modest number of loci. When this GSI approach was applied to green sturgeon encountered as bycatch in groundfish fisheries, to categorize them into DPS of origin, it showed that fish sampled in the southern fishery area, the Gulf of the Farallones, were almost entirely from the Southern DPS, whereas those in the northern area, surrounding the Columbia River plume, were a nearly

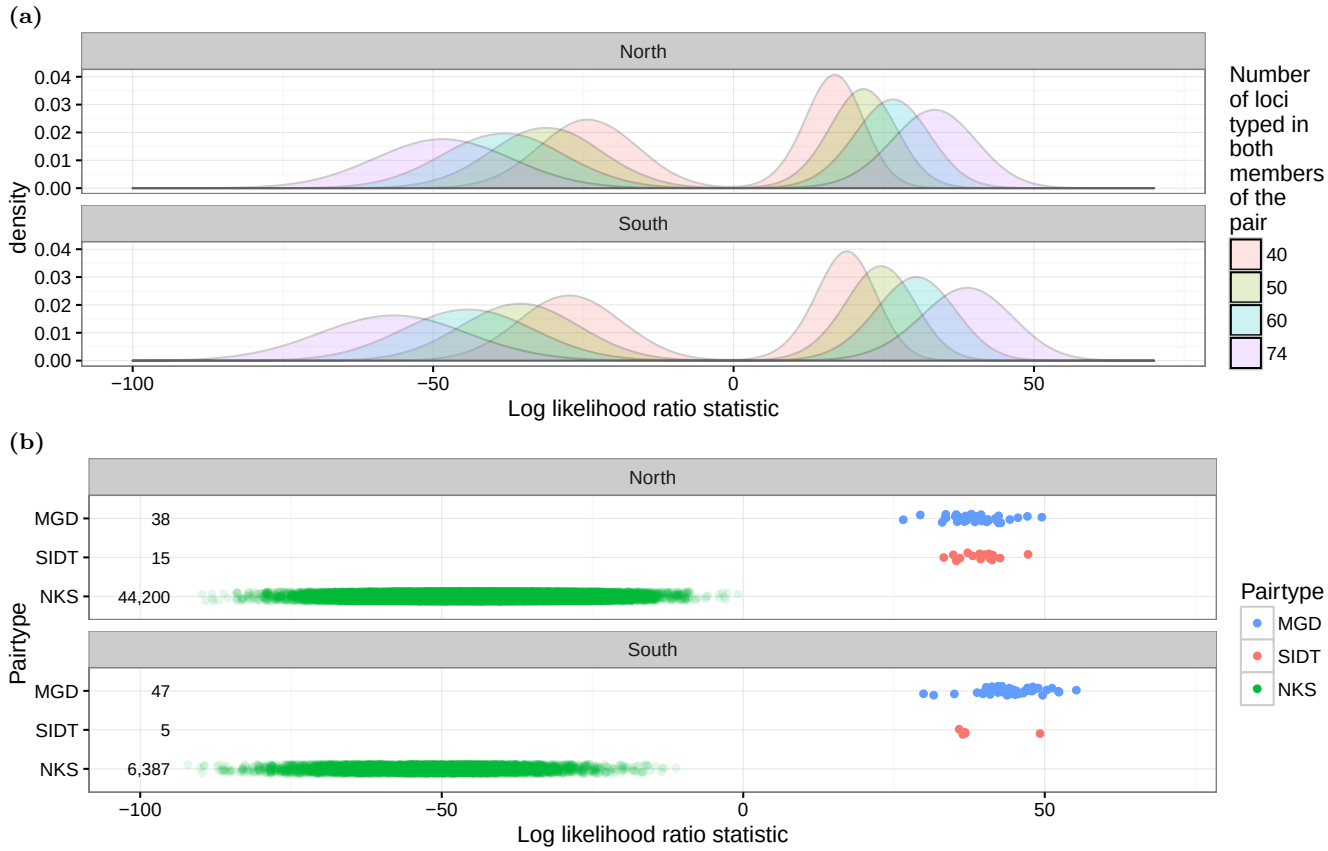

Fig. 7 Simulated (*a*) and observed (*b*) distributions of the log likelihood ratio statistic Λ_{ij} for identifying samples from the same individual. (*a*) The density curves on the left of the panel show the distribution of Λ_{ij} for unrelated pairs (*i.e.*, $K_{ij} = \mathrm{U}$) and the curves on the right (having mass at values > 0) show the distribution for samples from the same individual ($K_{ij} = \mathrm{S}$). Different colors denote numbers of loci typed in both members of the pair. Curves are density estimates of 10^5 simulated pairs for each number of loci. Simulations were done using genotype category frequencies in the Northern and Southern DPS as indicated. ϵ_ℓ was assumed to be 0.05 for each locus. (*b*) Values of Λ_{ij} observed for all pairs of individuals (from Northern or Southern DPS) with at least 40 loci typed on both members of the pair. Different colors denote pairs of three different types: **MGD**: "multiply genotyped DNA"—samples known to be the same sample genotyped on multiple chips; **SIDT**: "same individual, duplicate tissue"—samples treated in the lab as completely separate, though they were actually two different tissues sampled from the same individual; **NKS**: "not known to be self"—samples that are not known *a priori* to be of the same individual. Numbers to the left of the panel indicate the total number of pairs of each of the three types.

equal mix of fish from the two DPSs. This finding suggests that green sturgeon migration in the the coastal ocean is primarily north of their river basin of origin, as the southern fishery area is to the south of all of the Northern DPS spawning rivers and is dominated by Southern DPS fish. Additionally, we show that green sturgeon recently encountered in the Eel River were unambiguously from the Northern DPS. These fish were apparently reproducing (J. Strange & S. Kullman, pers. comm.), which, if accurate, may indicate a recolonization of the Eel River, as it is recognized as part of the Northern DPS, but its green sturgeon population was extirpated (or nearly so) in the last century (Adams et al. 2007). Finally, we show how these genetic assays allow the individual identification of green sturgeon, providing accurate identification of tissue samples from the same fish sampled multiple times, even when only a fraction of the assays are successfully genotyped. This bodes well for their use in forensic or other applications where DNA quality may be low.

As green sturgeon have experienced a recent whole genome duplication, the assays described here interrogate more than one genomic locus and do not provide typical genotype data, with variants from the maternal and paternal chromosomes individually discriminated, but rather provide data from four gene copies (two each from both the maternal and paternal chromosomes). It is not possible to discriminate the different gene copies and assign them to locus, so we developed an heuristic method for categorizing the relative fluorescence intensity data from these assays into "genotype categories." These categories are defined by clustering position on a two-dimensional graph of intensity of fluorescence of the two dyes from the two variant-specific probes. The

genotype categories are intended to represent the five potential genotype combinations of two identical loci with the same two alleles present. However, we do not know the exact underlying inheritance of these variants and how they correspond to these genotype categories and therefore consider each one as an individual "phenotypic" character. As such, this method does not produce data suitable for pedigree reconstruction or other such applications where an explicit genetic model is necessary. Additional molecular genetic data for green sturgeon and methodological research will be necessary to develop markers and analytical methods appropriate for applications with this species that must assume Mendelian segregation or otherwise employ an explicit genetic model. The general approach, however, of designing SNP genotyping assays and then using a combination of software-generated and post-processing categorization, facilitated by calibration across analytical runs, is applicable to any polyploid species and opens up new avenues for genetic identification of the myriad species with multiple genome copies.

One drawback of the inability to infer pedigree relationships with genotype category data is that we are unable to detect and account for siblings that may have been sampled together in some of our reference collections. As a consequence, the genotype categories inferred from the reference data might not represent as much variation from each DPS as one would achieve with entirely unrelated individuals (Waples and Anderson 2017). However, the fact that fish from reference, non-reference, and bycatch samples are all confidently inferred to be entirely (*i.e.*, $q = 0$ or $q = 1$) from one cluster or another in the *structure* analysis with $K = 2$ (Fig. 4) indicates that the presence of siblings (if any) amongst our reference samples has had no appreciable impact on our ability to identify fish from the two DPSs.

Other methods for calling the genotypes of tetraploid individuals using SNP assay intensity data have been recently introduced, most notably the method implemented in the R package FitTetra (Voorrips et al. 2011). We applied FitTetra to the green sturgeon data, but only 49 of the 96 SNPs were considered callable by the software, and upon visual inspection of the intensity data for the FitTetra-called genotypes, we did not have high confidence in their reproducibility. Futhermore, FitTetra appears not to have a good mechanism for using the genotypes of individuals that have been genotyped on multiple chips to guide the SNP selection and intensity-modeling process. This led us to develop our own procedure. It should be noted that the recommended procedure for using FitTetra involves a substantial decision-making effort by the user, including empirical adjustment of parameters and some hand curation. Our approach does the same, but with a more visual method, rather than using the underlying modeling framework of FitTetra. Because of the inherent variability in quantitative PCR-based assays, and the larger number of expected fluorescence clusters with polyploid data, we have implemented a graphical approach to account for chip-to-chip variability and ensure consistent calling of genotype categories. While not absolutely necessary, the use of such a consistency check should reduce calling errors due to stochastic variation in the relative fluorescence of the two assay dyes.

One potential drawback, resulting from our method's reliance on an initial set of genotypes to develop the chip calling methodology, occurs when trying to type individuals that carry genotypes that were not previously observed. If these genotypes fall outside of the range of previously seen variation, they might be recorded as missing genotypes rather than as new, extant, genotypes. If our method were attempted in a species with many previously undocumented and highly diverged populations, then it would be appropriate to consider recalibrating the calling methodology as new samples were analyzed. For identifying green sturgeon to DPS it does not appear that such a recalibration is currently necessary: even though some of the sturgeon taken in bycatch may have originated from populations that were not represented in our reference samples (*i.e.*, the Rogue River) or which had allele frequencies that had fluctuated temporally from those in our reference populations, there were no individuals that could not be identified to DPS without great certainty.

In summary, we described here a straightforward and reliable method for identifying green sturgeon from the two recognized DPSs. A panel of microsatellites has previously been described that provides useful resolution of the fish from the two DPSs (Israel et al. 2009). However, these novel SNP assays provide substantially better resolution than the microsatellites (Supplement 3) and are considerably less challenging to genotype and analyze. Additionally, they provide a means to unambiguously identify individuals that have been sampled more than once, which will be useful for the study of green sturgeon. In addition, the data strongly support the delineation of fish from the Klamath and Sacramento River basins into two separate DPSs. The finding of a predominance of southern DPS fish in the southern fishery area, as well as a substantial number of them in the Northern fishery area, indicates that green sturgeon bycatch in these fisheries is primarily impacting Southern DPS fish. Further research analyzing samples from fish known to have originated in the Rogue River, the only other basin with substantial green sturgeon reproduction, will be required to determine whether these

fish are appropriately grouped into the Northern DPS with Klamath River green sturgeon.

Acknowledgements We are grateful to Phaedra Doukakis, Melissa Neuman, and Susan Wang (NMFS West Coast Regional Office) for engaging us in this project with the goal of identifying green sturgeon bycatch, Toz Soto (Karuk Tribe) and Bill Poytress (US Fish and Wildlife Service) for providing samples of juvenile green sturgeon from the Klamath and Sacramento rivers, respectively, Vanessa Apkenas and Cassie Columbus (Southwest Fisheries Science Center) for assistance with laboratory analyses. Joshua Strange (Stillwater Sciences) and Stephen Kullman (Wiyot Tribe) provided samples from fish encountered in the Eel River, and observations regarding their reproductive status, and staff from the NMFS West Coast Fisheries Observer Program provided samples from fishery bycatch. Steven Lindley, Ethan Mora, and two anonymous referees provided helpful comments on the manuscript.

References

Adams PB, Grimes C, Hightower JE, Lindley ST, Moser ML, Parsley MJ (2007) Population status of North American green sturgeon, *Acipenser medirostris*. Environmental Biology of Fishes 79:339–356

Anderson EC, Waples RS, Kalinowski ST (2008) An improved method for predicting the accuracy of genetic stock identification. Canadian Journal of Fisheries and Aquatic Sciences 65:1475–1486

Andreakis N, Kooistra WHCF, Procaccini G (2009) High genetic diversity and connectivity in the polyploid invasive seaweed *Asparagopsis taxiformis* (*Bonnemaisoniales*) in the Mediterranean, explored with microsatellite alleles and multilocus genotypes. Molecular Ecology 18:212–226

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011) Stacks: building and genotyping loci de novo from short-read sequences. G3: Genes, Genomes, Genetics 1:171–182

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society, Series B 39:1–38

Havelka M, Kaspar V, Hulák M, Flajshans M (2011) Sturgeon genetics and cytogenetics: a review related to ploidy levels and interspecific hybridization. Folia Zoologica 60:93–103

Israel JA, Bando KJ, Anderson EC, May B (2009) Polyploid microsatellite data reveal stock complexity among estuarine North American green sturgeon (*Acipenser medirostris*). Canadian Journal of Fisheries and Aquatic Sciences 66:1491–1504

Ludwig A, Belfiore NM, Pitra C, Svirsky V, Jenneckens I (2001) Genome duplication events and functional reduction of ploidy levels in sturgeon (*Acipenser, Huso* and *Scaphirhynchus*). Genetics 158:1203–1215

Mable BK (2004) 'Why polyploidy is rarer in animals than in plants': myths and mechanisms. Biological Journal of the Linnean Society 82:453–66

NMFS (2015) Southern distinct population segment of the North American green sturgeon (*Acipenser medirostris*) 5-year review: summary and evaluation. Tech. rep., National Marine Fisheries Service, West Coast Region, Long Beach, CA, URL `https://swfsc.noaa.gov/publications/FED/00630.pdf`

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS ONE 7:e37,135

R Core Team (2015) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL `http://www.R-project.org/`

Rosenberg NA (2004) DISTRUCT: a program for the graphical display of population structure. Molecular Ecology Notes 4:137–138

Schreier A, Langness OP, Israel JA, Van Dyke E (2016) Further investigation of green sturgeon (*Acipenser medirostris*) distinct population segment composition in non-natal estuaries and preliminary evidence of Columbia River spawning. Environmental Biology of Fishes 99(12):1021–1032

Schreier AD, Mahardja B, May B (2013) Patterns of population structure vary across the range of the white sturgeon. Transactions of the American Fisheries Society 142:1273–1286

Voorrips RE, Gort G, Vosman B (2011) Genotype calling in tetraploid species from bi-allelic marker data using mixture models. BMC Bioinformatics 12:172

Waples RS, Anderson EC (2017) Purging putative siblings from population genetic data sets: a cautionary view. Molecular Ecology 26(5):1211–1224